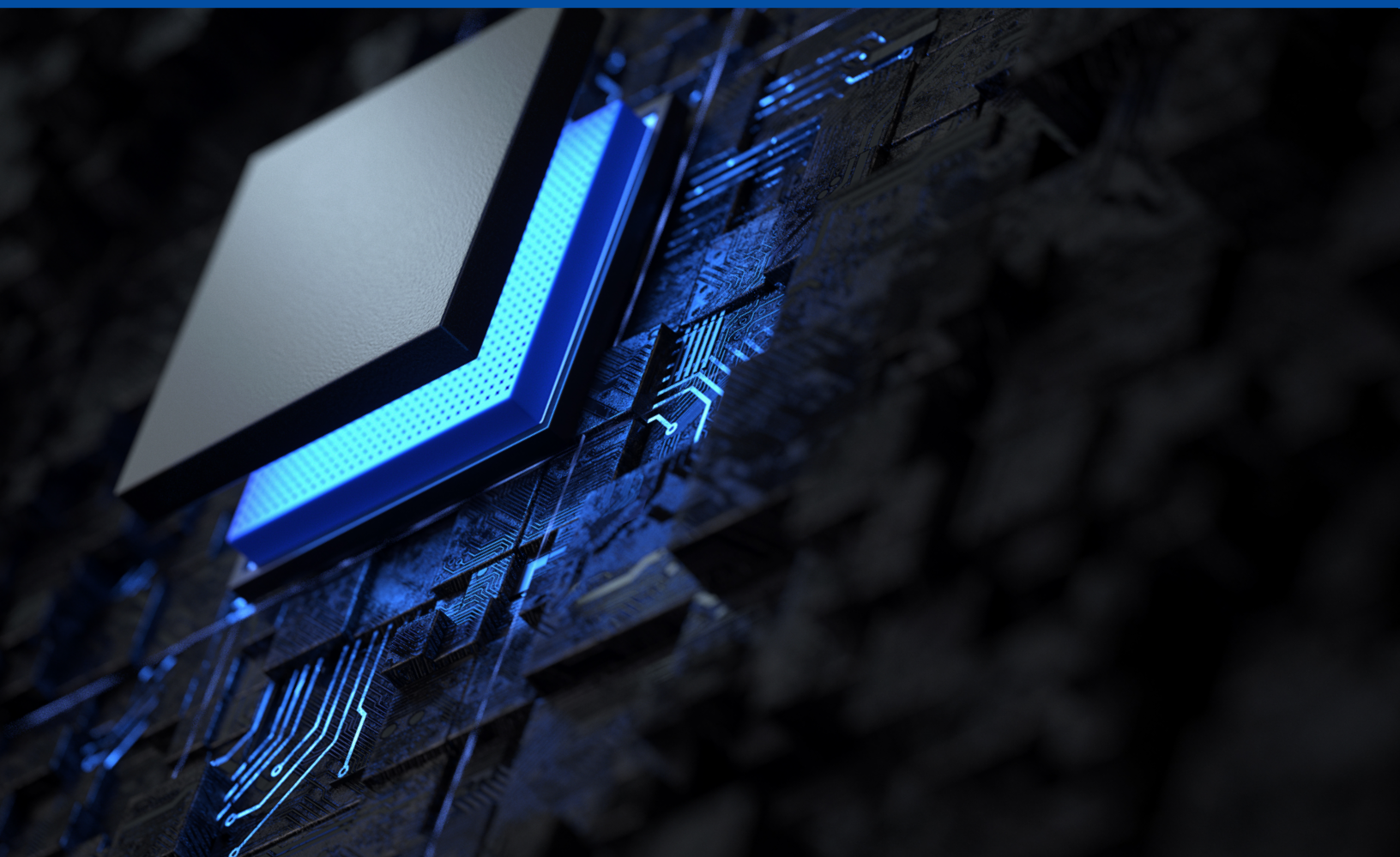
SAMSUNG

SVK (SSD Value Kit) Zero-ETL

White Paper

Authors: Ron Lee, Young Paik & Mayank Saxena



Samsung MSL Zero-ETL Solution

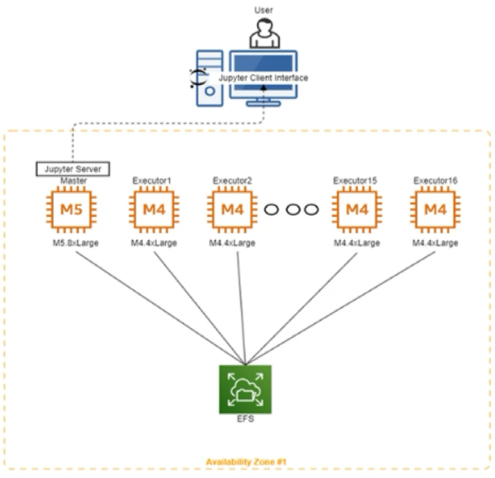
Samsung Zero-ETL Solution for Data Intensive AI Models

Samsung recently showcased their Zero-ETL feature that uses SSD Value Kit (SVK). In today's highly competitive business landscape, companies are willing to invest more in R&D to collect and analyze big data in real-time to gain insights into their operations, processes, and customers. However, with the increasing size of data sets and the sheer volume of data generated, companies are finding it challenging to analyze and extract useful data quickly.

As the Cloud leader, Amazon has recognized this and thus unveiled the concept of Zero-ETL (Etract, Transform, Load) at re:Invent 2022. Accordingly, Snowflake followed suit, launching hybrid tables and partnering with Salesforce to modernize data integration.
Zero-ETL is an approach that aims to minimize or eliminate the traditional ETL processes that are typically used to extract data from source systems, transform it into the desired format, and then load it into a data warehouse or data lake for analysis. In the context of Zero ETL, the data is often accessed in real-time or near real-time, eliminating the need for batch processing. This approach is becoming more popular with the advent of technologies like Change Data Capture (CDC) and streaming data pipelines.

As a world leader in memory storage technologies, Samsung collaborates with partners to innovate and create new enabling technologies, providing proven solutions for our customers. During a recent technology showcase, we presented our first generation Zero-ETL technology, which demonstrated a 40% faster speed compared to AWS EMR based ETL. This translates to hardware cost savings that are possible by reducing the number of analytics engines (such as Spark and Presto, among others) and eliminating the need to transfer bulk data from storage to analytics engines for batch processing. Below is the hardware cost gain achieved by offloading stochastic Data Generation and Filtering to the data source for Spark analytics application workload.

AWS EMR Topology



Samsung CoLo 5 (4+1) Node Set-Up

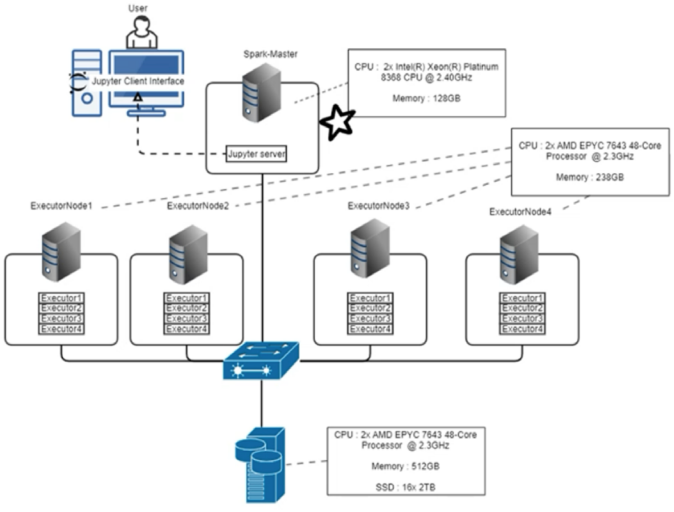


Figure 1

SAMSUNG

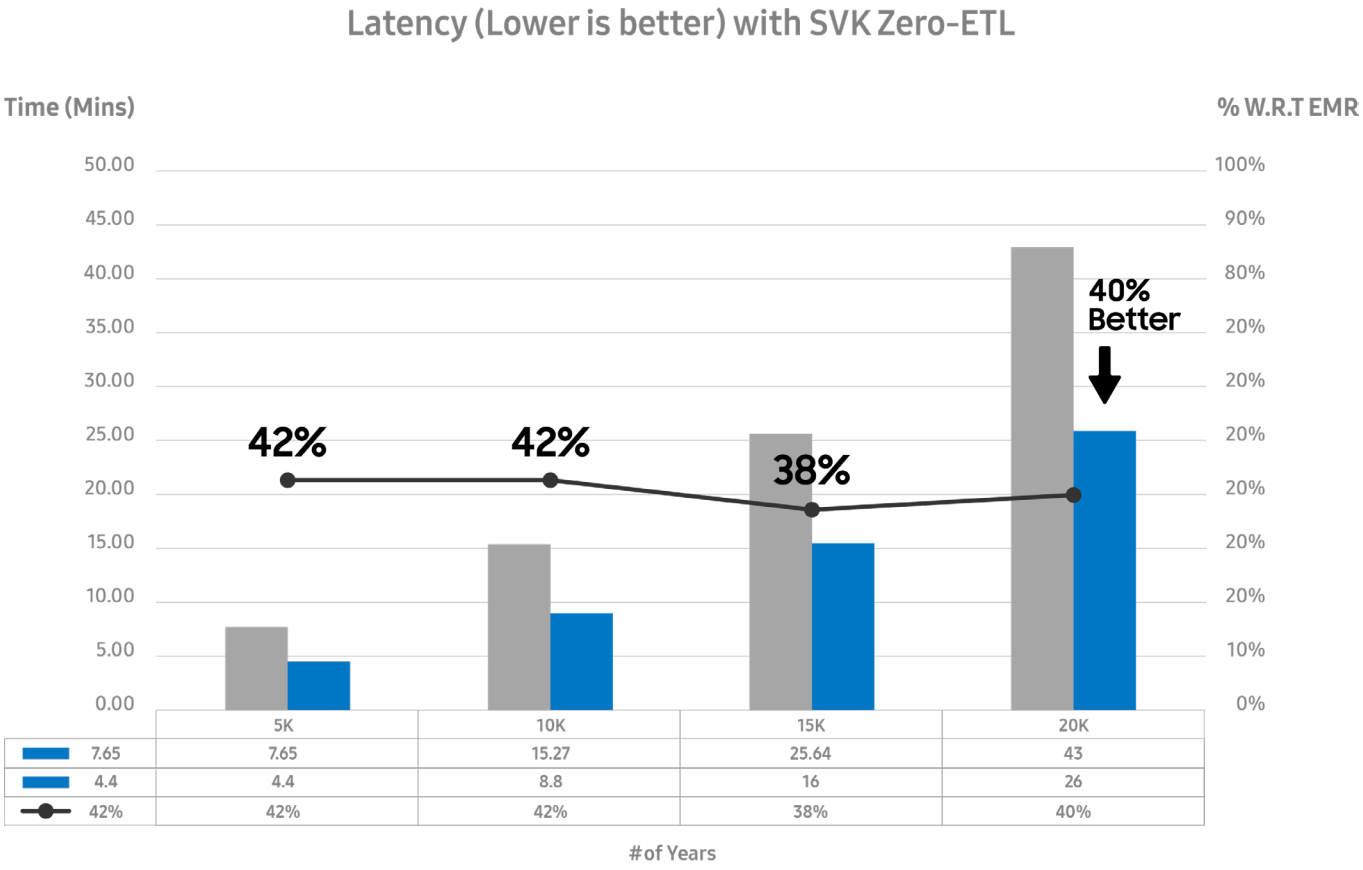


Figure 2

Samsung's Zero-ETL strategy is a multi-phased approach that will incrementally scale the data processing by moving much of the compute required to the data source. As shown in below diagram, the next Phase(3) will include offloading of compute functions to SVK Zero-ETL framework and further improve the hardware costs for this application.

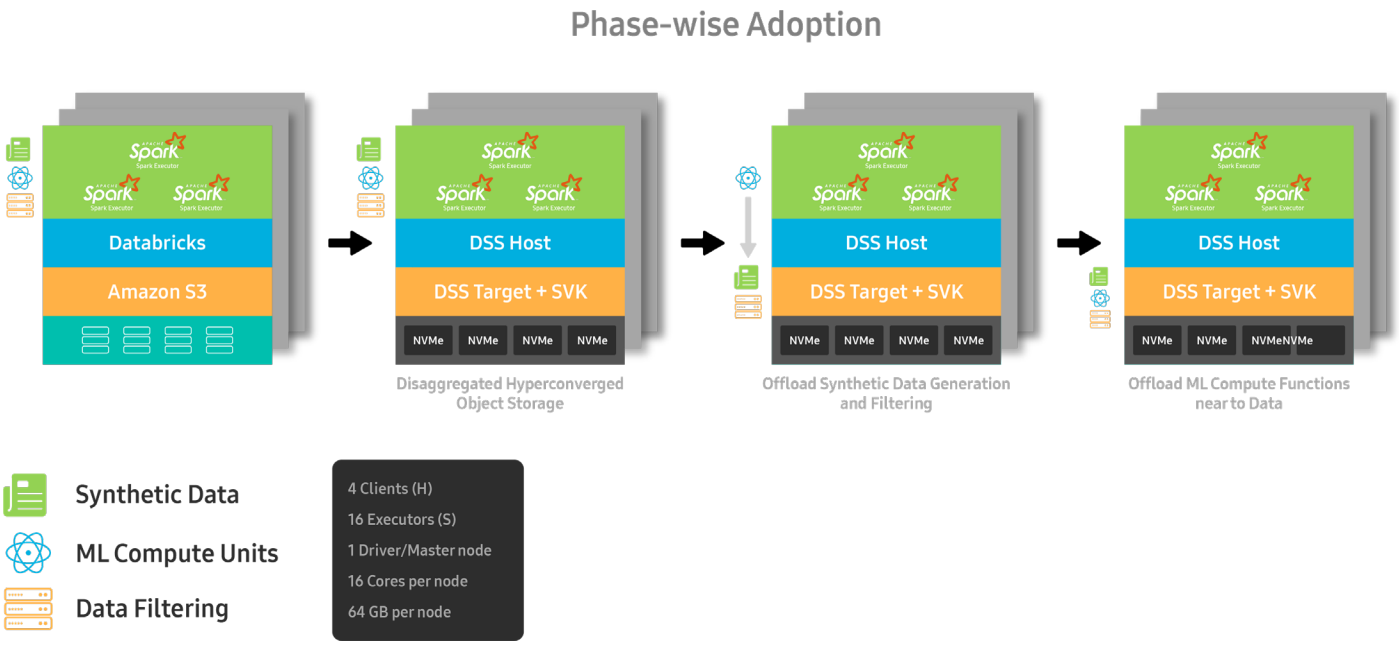


Figure 3

Samsung SVK(below diagram) provides the overall framework to support not only Zero-ETL but also Cluster Power Management and other vital services required to reap the maximum benefits from memory storage eco-system provided by Samsung, with ease of adoption i.e. simple REST API based integration to AI applications directly or to Software Defined Storage host.

In addition, Jupyter Notebook can be seamlessly used with Spark and SVK to provide simple user interface to quickly and visually run analytics. By integrating Jupyter Notebook with Spark+SVK, data scientists can tap the power of both Spark Cluster and SVK framework. As an example, PySpark can be configured with the following options to invoke Jupyter Notebook instances.

```
export PYSPARK_DRIVER_PYTHON='jupyter'
export PYSPARK_DRIVER_PYTHON_OPTS='notebook --no-browser --port=8889'
```

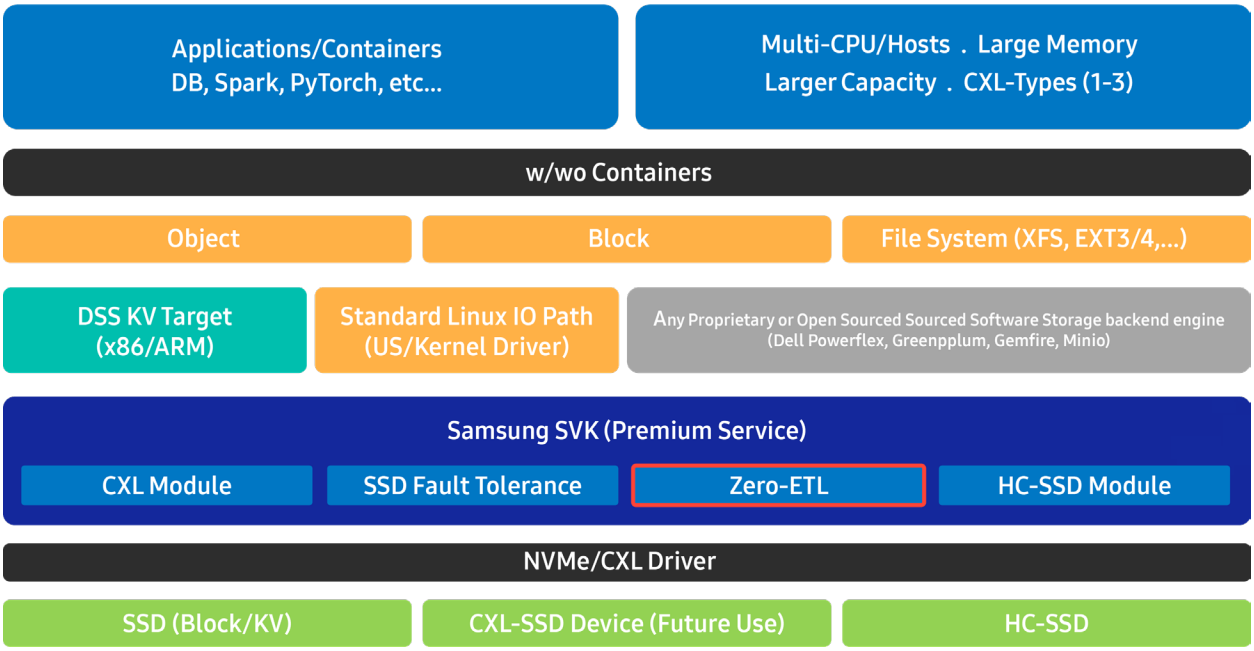


Figure 4

Analytics applications that wants to utilize Samsung's SVK Zero-ETL technology can easily integrate their data pipeline to leverage the provided SVK Shared Library to offload compute operations as shown below.

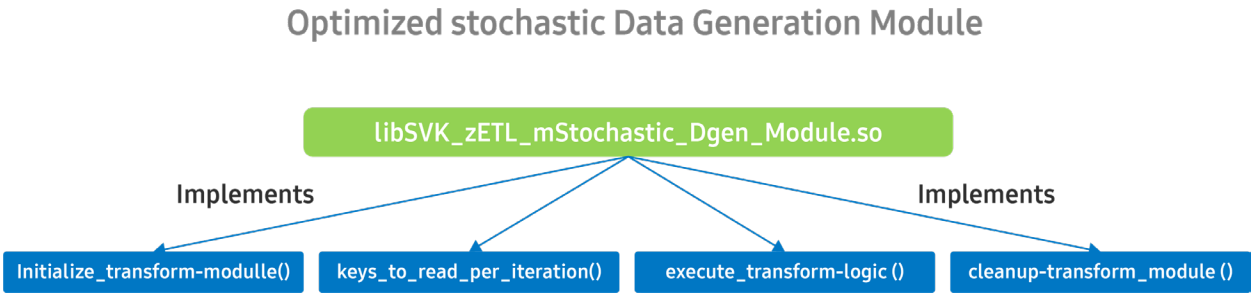


Figure 5

The data source or DSS (Samsung Open Sourced S3 Storage Node(s)) would implement the transform functions shown above either with resident server CPU or in the future utilize advanced CXL-SSD technology to handle the compute operations. By handling vast majority of compute operations for transformation to the DSS or the data source, the volume of data transferred for ETL will be drastically reduced and thus provide tremendous Cloud OpeEx savings since there is a major cost to moving data in the Cloud. In addition, customers will be able to quickly digest the real-time data as they arrive by offloading and distributing this task to the data source where it can be processed quickly and also be shared by multiple analytics applications in parallel. Below shows potential cost saving with Samsung Zero-ETL for typical models, with 500 runs, compare to the ETL in cloud.

Data Size	Application Compute OpEx Savings ($M)	Data Tranfer OpEx Savings ($M)	Total Savings ($M)
100 TB	0.4	0.7	1.1
0.5 PB	2.2	3.3	5.5
1 PB	4.5	6.7	11.2

Samsung Zero-ETL feature is first of its kind for hybrid cloud, private clould and multicloud deployments. It also unviled how any customer preferred object storage can easily intergrate this feature as part of their toolkit utilizing Samsung SSD Value Kit. In summary here are the key benefits:

- Lower compute and data transfer cost
- Multi-Cloud readiness
- Retain ownership of data

For more information, please visit https://samsungmsl.com/dfs/

For more information

For more information about the Samsung Semiconductor products, visit semiconductor.samsung.com.

About Samsung Electronics Co., Ltd.

Samsung Electronics Co. Ltd inspires the world and shapes the future with transformative ideas and technologies. The company is redefining the worlds of TVs, smartphones, wearable devices, tablets, digital appliances, network systems, memory, system LSI and LED solution. For the latest news, please visit the Samsung Newsroom at news.samsung.com.

Samsung Electronics Co., Ltd.
129 Samsung-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do 16677, Korea **www.samsung.com** 1995-2021

SAMSUNG